Research Article

Machine Learning Algorithms for Predicting Energy Consumption in Educational Buildings

Khaoula Elhabyb,1 Amine Baina,1 Mostafa Bellafkih,1 and Ahmed Farouk Deifalla2

1National Institute of Posts and Telecommunications (INPT), Rabat, Morocco
2Future University Cairo in Egypt, Cairo, Egypt

Correspondence should be addressed to Ahmed Farouk Deifalla; ahmed.deifalla@fue.edu.eg

Received 14 December 2023; Revised 19 March 2024; Accepted 26 March 2024; Published 13 May 2024

Academic Editor: Saleh N. Al-Saadi

Copyright © 2024 Khaoula Elhabyb et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In the past few years, there has been a notable interest in the application of machine learning methods to enhance energy efficiency in the smart building industry. The paper discusses the use of machine learning in smart buildings to improve energy efficiency by analyzing data on energy usage, occupancy patterns, and environmental conditions. The study focuses on implementing and evaluating energy consumption prediction models using algorithms like long short-term memory (LSTM), random forest, and gradient boosting regressor. Real-life case studies on educational buildings are conducted to assess the practical applicability of these models. The data is rigorously analyzed and preprocessed, and performance metrics such as root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) are used to compare the effectiveness of the algorithms. The results highlight the importance of tailoring predictive models to the specific characteristics of each building’s energy consumption.

1. Introduction

Artificial intelligence is rapidly being integrated into various industries, such as healthcare, finance, and smart grids. Among these human-centric applications, the use of AI in smart buildings has attracted significant attention from a large community [1]. Smart buildings, which have been a subject of research since the 1980s, utilize advanced technology, data analytics, and automation systems to optimize operations, enhance occupant comfort and productivity, and reduce costs and energy consumption [2]. These buildings incorporate sensors, devices, and control systems to monitor lighting, HVAC systems, security, and access controls. Real-time data on occupancy, temperature, air quality, and energy use can be analyzed to identify optimization opportunities. The primary aim is to create an efficient, comfortable, and sustainable environment for residents while reducing costs and ecological impact.

The smart building industry is experiencing significant growth as society becomes more connected and digital. According to statistics from MarketsandMarkets [3], the industry is projected to expand at a compound annual growth rate (CAGR) of 10.5% between 2020 and 2025, reaching a value of $108.9 billion. This growth is driven by factors such as increased energy usage and expenses, advancements in machine learning and the Internet of Things (IoT), the push for net zero energy buildings, and regulatory changes that encourage the adoption of smart building systems and services. Figure 1 presents the forecasted global market side from 2020 to 2030. Expanding on the findings of the Zion Marketing research study [4], it reveals the market value of 40,760 million in 2016, with projections of a substantial growth trajectory to 61,900 million by 2024, with a CAGR exceeding 34%. This indicates a rapid expansion within the market, indicating robust trends and significant economic development during the study period.

The AI sector being discussed is experiencing significant growth due to the integration of the Internet of Things (IoT) and machine learning (ML). IoT sensors collect data about buildings and occupants, such as temperature, humidity, occupancy, and electricity consumption.
This data is centralized for optimizing building operations, improving resident comfort, and reducing energy usage. ML, on the other hand, is a powerful tool for processing large amounts of data from various sources. It analyzes this data to identify patterns and predict future events, such as equipment failures, enabling preventative maintenance [5].

The American Council for Energy-Efficient Economy (ACEEE) [6] suggests that commercial buildings can significantly reduce their energy bills by up to 30% by implementing energy-efficient technologies such as smart thermostats and controlled lighting. The US Department of Energy [7] reports that commercial buildings account for a significant portion of total energy consumption and greenhouse gas emissions in the US. This highlights the importance of buildings that can predict energy consumption and plan efficiently to reduce energy usage. Intel research [8] indicates also that energy consumption prediction has the potential to achieve operational cost savings, staff productivity gains, and energy usage reductions. Given these findings, the primary emphasis will be on forecasting the energy usage of smart buildings, with a specific focus on educational facilities, which will be analyzed for the first time. Understanding and predicting energy consumption in educational environments are paramount for optimizing resource allocation, implementing effective efficiency measures, and establishing sustainable and cost-effective operational procedures [9]. By focusing on this sector, valuable insights can be gained to inform strategies for enhancing energy efficiency and sustainability in educational buildings, ultimately contributing to improved resource management and environmental conservation efforts.

The research concentrates on energy management within smart buildings, aiming to forecast power consumption through three distinct approaches: a traditional statistical approach employing the random forest algorithm, a deep learning approach utilizing long short-term memory (LSTM), and a hybrid approach leveraging the gradient boosting regressor algorithm. These three techniques were chosen to investigate a research gap regarding to the majority of data-driven methodologies. While significant progress has been made in this area, limited attention has been given to utilizing streaming and temporal data for forecasting buildings’ energy demand. This gap will be addressed through the utilization of real historical electricity data. The used data is analyzed to evaluate model performance and accuracy, aiming to identify the most effective approach for smart building energy management. The research is aimed at optimizing forecasting techniques through rigorous comparative analysis, leveraging the strengths of LSTM, RF, and GBR models. The study highlights the importance of advanced machine learning in shaping smart building strategies and is aimed at enhancing sustainability and efficiency in energy usage. Insights from this research will inform future advancements in energy management practices for sustainable development. The article is structured into several delineated sections, each serving a specific purpose:

(i) Introduction: This section introduces the application of AI within the smart building sector, setting the context for the study

(ii) Literature analysis: Here, a comparative examination of various ML algorithms used for energy prediction in smart building systems is provided, drawing insights from existing research

(iii) Methodology: This section outlines the systematic approach adopted in the study, encompassing data analytics, model development, and model evaluation processes

(iv) Results and discussions: Findings obtained from the methodology are presented, followed by a comparative analysis that juxtaposes these results with prior research initiatives

(v) Conclusion: This section synthesizes the results and provides conclusions, offering perspectives on the implications of the study’s findings for the field of smart building energy management

2. Literature Review

A recent study conducted by the International Energy Agency [10] has revealed concerning levels of energy
consumption in buildings. The study found that buildings are responsible for a significant portion of electricity consumption and overall energy consumption in urban areas. Buildings account for 72% of total electricity consumption and 38% of average energy consumption in urban areas. Additionally, buildings contribute to almost 40% of total carbon dioxide pollution in urban areas. A smart building is a modern infrastructure that incorporates automated control systems and uses data to improve the building’s performance and occupants’ comfort. Figure 2 presents the smart building functionalities and its most important axis of work.

The top technology companies are currently prioritizing IoT (Internet of Things) and AI (artificial intelligence). The future of building innovation is expected to focus on achieving maximum energy efficiency, and this challenge can be addressed by integrating AI-powered systems like machine learning (ML) and deep learning. ML systems continuously improve themselves, leading to advancements in various AI research areas [12]. ML involves algorithms that allow them to respond to inputs from their environment and identify nonlinear connections in complicated or uncertain systems. ML is divided into four major categories based on the type of learning task they manage: supervised learning, unsupervised learning, semisupervised learning, and reinforcement learning.

(i) Supervised learning is a method of developing a machine learning model by using a labeled data set. In this process, each data point in the set is associated with a known intended output. The model is trained to predict the output

(ii) Unsupervised learning: in contrast to traditional supervised learning, developing a model on an unlabeled data set involves working with data where the target outputs are unknown. In this scenario, the model is not explicitly instructed on what to search for but instead learns

(iii) Semisupervised learning is a learning approach that combines supervised and unsupervised learning. In this approach, the model is trained using a data set that is partly labeled, meaning that some of the data points have known labels

(iv) Reinforcement learning where a model is trained to make a series of decisions in a changing environment. The model learns through trial and error, receiving feedback in the form of rewards or costs

Energy consumption prediction is a valuable technique that involves forecasting the amount of energy a system or device will use within a specific time frame. This technique serves various purposes, such as optimizing energy usage, predicting future energy demands, and identifying potential inefficiencies in energy consumption. To predict energy consumption, different methods can be employed, including statistical models, machine learning algorithms, and physics-based models. The choice of technique depends on factors such as data availability, system complexity, and the desired level of accuracy. In this particular case, the focus is on utilizing machine learning algorithms to predict energy consumption by leveraging historical data and other relevant factors.

The quality and relevance of the data used in machine learning algorithms greatly influence their performance. In a study conducted by Ahajjam et al. [13] on Moroccan Buildings’ Electricity Consumption Data Set, electricity consumption was categorized into three types: whole premises (WP), individual loads (IL), and circuit-level (CL) data.

(1) Labeled WP: Labeled whole premises (WP) consumption data refers to electricity usage data collected from 13 households in the MORED data set.
This data is valuable as it includes not only the raw electricity consumption measurements but also additional information that can assist in analyzing, modeling, and comprehending the patterns of electricity usage in different households.

2. Labeled IL: Ground-truth electricity consumption refers to the electricity consumption data of individual loads (IL) that have been labeled or annotated with accurate information. This involves recording and labeling the operational states of specific loads, such as refrigerators or air conditioners when they are turned on or off at specific times. Having this ground-truth information is valuable for researchers and analysts as it allows for accurate load disaggregation, energy management, and appliance recognition.

3. CL: Measurements in the context of energy refer to the circuit-level energy measurements obtained from the electrical mains of a premises. These measurements provide information about the overall energy consumption of a circuit and can be used to understand the energy consumption of a group of loads.

The current work focuses on three educational buildings located at Down Town University. Further information about these buildings will be provided next. The subsequent section presents a literature review on energy consumption forecasting in various buildings using multiple machine learning algorithms.

2.1. Traditional Machine Learning Approach. ML algorithms have been utilized to tackle the primary challenges of physics-driven methods in load prediction. For instance, Somu et al. [14] developed eDemand, a new building energy use forecasting model, using long short-term memory networks and an improved sine cosine optimization algorithm, and as a result, the model outperformed previous state-of-the-art models in real-time energy load prediction. Next, Suranata et al. [15] focused on predicting energy consumption in kitchens. They used a feature engineering technique and a short-term memory (LSTM) model. Principal component analysis (PCA) was applied to extract important features, and the LSTM model was used on two tables. In addition, Shapi et al. [16] developed a prediction model for energy demand making use of the Microsoft Azure cloud-based machine learning framework. The methodology of the prediction model is provided using three distinct techniques, including support vector machine, artificial neural network, and k-nearest neighbors. The study focuses on real-world applications in Malaysia, with two tenants from an industrial structure chosen as case studies. The experimental findings show that each tenant’s energy consumption has a particular distribution pattern, and the suggested model can accurately estimate energy consumption for each renter. To forecast daily energy consumption based on weather data, Faq et al. [17] developed a new energy usage prediction technique for institutional buildings using long short-term memory (LSTM). The model, trained using Malaysian Meteorological Department weather forecasting data, outperformed support vector regression (SVR) and Gaussian process regression (GPR) with the best RMSE scores. The dropout method reduces overfitting, and Shapley’s additive explanation is used for feature analysis. Accurate energy consumption estimates can help detect and diagnose system faults in buildings, aiding in energy policy implementation. Further, Kawahara et al. [18] explore the application of various machine learning models to predict voltage in lithium-ion batteries. The study includes algorithms such as support vector regression, Gaussian process regression, and multilayer perceptron. The hyperparameters of each model were optimized using 5-fold cross-validation on training data. The data set used consists of both simulation data, generated by combining driving patterns and applying an electrochemical model, and experimental data. The performance of the ML models was evaluated using both simulation and experimental data, with different data sets created to simulate variations in state of charge distribution.

2.2. Deep Learning and Hybrid Approaches. Additionally, various networks integrate multiple techniques to devise data-driven approaches. These integrated mechanisms are commonly referred to as hybrid networks. For example, Mohammed et al. [19] focus on the application of an intelligent control algorithm in HVAC systems to enhance energy efficiency and thermal comfort. The authors propose integrating SCADA systems with an intelligent building management system to optimize heat transmission coefficients and air temperature values. Genetic algorithms are employed to maintain user comfort while minimizing energy consumption. Similar to [19], Aurna et al. [20] compare the performance of ARIMA and Holt-Winters models in predicting energy consumption data in Ohio and Kentucky. The study finds that the Holt-Winters model is more accurate and effective for long-term forecasting. The authors recommend further research to consider other parameters, and environmental factors, and explore hybrid models for better short-term load forecasting. Next, Ferdoush et al. [21] developed a hybrid forecasting model for time series electrical load data. The model combines random forest and bidirectional long short-term memory methods and was tested on a 36-month Bangladeshi electricity consumption data set. The results showed that the hybrid model outperformed standard models in terms of accuracy. The study emphasizes the effectiveness of the hybrid machine learning approach in improving short-term load forecasting accuracy in the dynamic electric industry. In their study, He and Tsang [22] developed a hybrid network combining long short-term memory (LSTM) and improved complete ensemble empirical mode decomposition with adaptive noise (iCEEMDAN) to optimize electricity consumption. They divided the initial power consumption data into patterns using iCEEMDAN and used Bayesian-optimized LSTM to forecast each mode independently. In the same direction, Jin et al. [23] proposed an attention-based encoder-decoder network with Bayesian optimization for short-term electrical load forecasting, using a gated recurrent unit recurrent neural network for time series data modeling and a temporal attention layer for improved prediction accuracy and
3. Methodology

This research predicts power usage in three buildings of a private research university using a data set collected from January 2020 to January 2023. The university is known for its excellence in education and research across various disciplines. The buildings under study (referred to as CLAS, NHAI, and Cronkite) are all part of the same institution and serve distinct functions. Building CLAS, an abbreviation of Center of Law and Society, mainly consists of an amphitheater and offices, and building NHAI, which means Nursing and Health Innovation, consists of offices and laboratories. In contrast, Cronkite consists of classrooms and seminar halls.

The buildings are equipped with IoT sensors connected to power intel sockets, and the collected data is sorted on an open-source website server [32]. The prediction method will use three machine learning algorithms: long short-term memory (LSTM), random forest (RF), and gradient boosting regressor (GBR). The data will be analyzed and prepared before being used to train and test the models.

The methodology for forecasting energy consumption will be divided into three sections:

(1) Data analysis involves evaluating raw data to understand patterns and characteristics of electrical power consumption data.

(2) Model training trains machine learning models, using past data to identify patterns and correlations between input characteristics and day power use.

(3) Model test models evaluation using validation metrics to assess their performance and accuracy.

3.1. Data Analysis

3.1.1. Data Preparation. This study focuses on the process of data preparation in machine learning, which is time-consuming and computationally challenging due to the presence of missing values and uneven value scales between features. The data was prepared using two techniques: imputation of missing data and standardization. The imputation procedure was carried out using the probabilistic principal component analysis (PPCA) approach, a maximum likelihood estimate-based technique that estimates missing values using the expectation-maximization (EM) algorithm. This method is developed from the principal component analysis (PCA) method, which is used for data compression or dimensionality reduction. The resulting cleaned data was then subjected to standardization, also known as Z-score normalization, to ensure an even distribution of the data above and below the mean value as shown in equation (3):

\[ x_{\text{standardized}} = \frac{x - \mu}{\sigma}, \]  

where \( \mu \) represents the mean, \( \sigma \) denotes the standard deviation, and \( x \) is the original data points.

3.1.2. Data Normality Analysis. This research conducted a normality test on each renter’s data set to determine its distribution. This test is crucial for model construction and is especially important for larger sample sizes. Understanding the data set distribution can provide valuable insights into the prediction outcome. Kurtosis measures distribution peaks,
<table>
<thead>
<tr>
<th>Authors</th>
<th>Algorithm</th>
<th>Data set</th>
<th>Performance evaluation</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Somu et al.</td>
<td>(i) ARIMA</td>
<td>(i) The KReST power consumption data set is sourced from the Indian Institute of Technology (IIT) in Mumbai, India.</td>
<td>(i) ARIMA: MAE 0.3479; MAPE: 0.1661; RMSE: 0.4076.</td>
<td>(i) Improved forecasting accuracy</td>
<td>(i) Sensitivity to initialization (ii) Convergence speed</td>
</tr>
<tr>
<td></td>
<td>(ii) Genetic algorithm-LSTM</td>
<td></td>
<td>(ii) Genetic algorithm-LSTM: MAE 0.1804; MAPE: 5.9745; MSE 0.0432; RMSE: 0.2073.</td>
<td>(ii) Improved forecasting accuracy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(iii) Sine cosine optimization</td>
<td></td>
<td>(iii) ISCOA-LSTM: MAE 0.0818; MAPE 4.9688; MSE 0.0135; RMSE 0.1164.</td>
<td>(iii) Real-world applicability</td>
<td></td>
</tr>
<tr>
<td></td>
<td>algorithm-LSTM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suranata et al.</td>
<td>(i) Long short-term memory</td>
<td>(i) NL</td>
<td>(i) RMSE 62.013; MAE 26.982; MAPE 12.876.</td>
<td>(i) The ability to effectively predict energy consumption patterns in time series data.</td>
<td>(i) Time-consuming training</td>
</tr>
<tr>
<td>Ferdoush et al.</td>
<td>(i) LSTM</td>
<td>(i) Bangladesh Power Development Board covered 36 months.</td>
<td>(i) LSTM: MSE = 0.4776; RMSE = 0.691; MAE = 0.5578; MAPE = 148.7.</td>
<td>(i) Stable learning characteristics</td>
<td>(i) The hybrid model may require specific data to utilize the strengths of random forest and bidirectional LSTM effectively.</td>
</tr>
<tr>
<td></td>
<td>(ii) RF-bi-LSTM hybrid model</td>
<td></td>
<td>(ii) Bi-LSTM: MSE = 0.2943; RMSE = 0.5425; MAE = 0.4317; MAPE = 194.80.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(iii) Bidirectional long short-term memory (LSTM)</td>
<td></td>
<td>(iii) RF-bi-LSTM: MSE = 0.1673; RMSE = 0.4090; MAE = 0.3070; MAPE = 193.49.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yaqing et al.</td>
<td>(i) EMD-BO-LSTM</td>
<td>(i) Real power consumption data of a university campus for 12 months</td>
<td>(i) EMD-BO-LSTM: MAE = 155.77; RMSE = 203.4; MAPE = 10.41%.</td>
<td>(i) Adaptability and efficiency</td>
<td>(i) NL</td>
</tr>
<tr>
<td></td>
<td>(ii) iCEEMDAN</td>
<td></td>
<td>(ii) iCEEMDAN-BO-LSTM: MAE = 40.81; RMSE = 59.68; MAPE = 2.5563%; R2 = 0.986.</td>
<td>(ii) Enhanced prediction accuracy</td>
<td></td>
</tr>
<tr>
<td>Ndife et al.</td>
<td>(i) ConvLSTM encoder-decoder</td>
<td>(i) Two million measurements were gathered over 47 months from a residential location in Sceaux, France.</td>
<td>(i) RMSE on the model: 358 kWh; RMSE on the persistence model: 465 kWh; RMSE on model A: 530 kWh; RMSE on model B: 450.5 kWh</td>
<td>(i) Improved forecast accuracy</td>
<td>(i) Model complexity</td>
</tr>
<tr>
<td>Duong et al.</td>
<td>(i) Multiple layer perceptron</td>
<td>(i) 215 data points on the power consumption and on/off status of electrical devices, in Vietnam.</td>
<td>(i) RMSE 10.468; (ii) MAE 21.563</td>
<td>(ii) Suitable for low-powered devices</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(ii) Efficient training and prediction time</td>
<td></td>
</tr>
<tr>
<td>Faq et al.</td>
<td>(i) LSTM</td>
<td>(i) Daily data from 2018 to 2021, from the Malaysian Meteorological Department.</td>
<td>(i) LSTM: RMSE = 165.20; MAE = 572.54.</td>
<td>(i) Accurate prediction of building energy consumption</td>
<td>(i) Requires a significant amount of historical data to create an accurate model</td>
</tr>
<tr>
<td></td>
<td>(ii) LSTM-RNN</td>
<td></td>
<td>(ii) LSTM-RNN: RMSE = 263.14; MAE = 353.38.</td>
<td>(ii) Improved energy efficiency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(iii) CNN-LSTM</td>
<td></td>
<td>(iii) CNN-LSTM: RMSE = 692.14; MAE = 1134.1.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bhol et al.</td>
<td>(i) ARIMA</td>
<td>(i) Laboratory-operated critical loads over three months.</td>
<td>(i) HW-GFPA: MBE = 0.42 for validation, 0.43 for test RMSE = 0.80.</td>
<td>(i) Scalability</td>
<td>(ii) Optimal hyperparameter identification</td>
</tr>
<tr>
<td></td>
<td>(ii) Holt-Winters flower pollination algorithm</td>
<td></td>
<td>(ii) ARIMA: MBE = 0.073 for validation, 0.016 for test RMSE = 0.183</td>
<td>(ii) Optimal hyperparameter identification</td>
<td>(i) Sensitivity to kernel selection</td>
</tr>
</tbody>
</table>
while skewness measures irregular probability distribution around the mean value [33]. Equations (2) and (3) provide formulas for skewness and kurtosis, which are essential for understanding the data set distribution and its impact on the prediction outcome.

\[
\text{Skewness} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^3}{N - 1 + \sigma^3},
\]

\[
\text{Kurtosis} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^4}{\sigma^4},
\]

where \(n\) is the number of data points in the collection, \(x_i\) is the individual data points within the sample, and \(\bar{x}\) is the sample mean.

### 3.1.3. Feature Selection

Feature engineering is a crucial aspect of machine learning, involving the creation of meaningful data representations to enhance model performance. It involves careful selection, transformation, and creation of features that capture relevant information from raw data, enhancing predictive accuracy and interoperability. Techniques like principal component analysis, domain knowledge extraction, and creative data manipulation help models extract patterns and make accurate predictions, bridging the gap between raw data and actionable insights.

As previously stated, our data set comprises 27 features detailing the characteristics of the selected buildings. To ensure optimal input for our predictive model, we employed a feature engineering approach leveraging a tree-based model, specifically the random forest algorithm.

### 3.2. Model Development

This study uses supervised machine learning to predict energy usage using data prepared and trained in two groups. The model employs regressive prediction using random forest, LSTM, and gradient boosting regressor. The process from data collection to model generation is depicted in Figure 3.

#### 3.2.1. Random Forest

A random forest regressor is a machine learning method that combines multiple decision trees to create a predictive model for regression tasks. Each tree is constructed using a randomly selected subset of training data and features with \(H(x; \theta_k), k = 1, \ldots, K\) where \(x\) represents the observed input (covariate) vector of length \(p\) with associated random vector \(X\). During prediction, the regressor aggregates predictions from all trees to generate the final output, typically the average of the individual three predictions \(h(x) = (1/k) \sum_{k=1}^{K} h(x; \theta_k)\) [34]. This method is commonly used for pattern identification and prediction due to its ability to learn complicated behavior. Consequently, it is the best choice for constructing the prediction model in the present study. In Figure 4, we present a flow chart of the random forest algorithm.

#### 3.2.2. Long Short-Term Memory

Sepp Hochreiter and Jürgen Schmidhuber introduced long short-term memory (LSTM) in 1997 as an advanced application of recurrent neural networks. LSTM is effective in processing and predicting time series data with varying durations. It captures long-term relationships, handles variable-length sequences, and recalls previous data, making it useful for energy consumption prediction [35]. The LSTM model structure consists of three layers: input, LSTM unit, and output. The mathematical equations used in LSTM include the forget gate, input gate, output gate, and cell state. The following are the equations utilized in LSTM:

\[
i_t = \sigma(W_f \cdot [ht_{t-1}, x_t] + b_f),
\]

\[
C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t,
\]

\[
o_t = \sigma(W_o \cdot [ht_{t-1}, x_t] + b_o)h_t = o_t \cdot \tan(h(C_t),
\]

where \(x_t\) is the input at the step \(t\); \(i_t, f_t,\) and \(o_t\) are the input, forget, and output vectors; \(g_t\) is the candidate activation vector, and \(c_t\) is the cell state at time \(t\).

The LSTM algorithm is a powerful tool for collecting and transmitting information across long sequences. It is commonly used in applications such as audio recognition, natural language processing, and time series analysis. Based on previous research and the availability of a time series data set, LSTM is chosen as the algorithm for predicting energy with high precision. Figure 5 presents a flowchart of LSTM.

#### 3.2.3. Gradient Boosting Regressor

The gradient boosting approach is an iterative method that combines weak learners to create a strong learner by focusing on errors at every step. It is aimed at decreasing the loss function by matching weak learner models to the loss function [36]. The squared error function is often used to estimate the approximation function, which is then used to find the ideal settings for weak learners. The gradient boosting regressor’s mathematical equation is as follows:

\[
y_i = F(x_i) + \sum_{m=1}^{M} \gamma_m h_m(x_i),
\]

where \(y_i\) is the predicted target, \(x_i\) is the input features, \(F(x_i)\) is the ensemble model prediction, \(M\) is the weak model, \(\gamma_m\) is the learning rate, and \(h_m(x_i)\) is the prediction by \(m\)-th weak model. The current research utilized gradient boosting due to its robust predictive performance, ability to capture complex data linkages and nonlinear patterns, and flexibility and customization capabilities. Figure 6 depicts the gradient boost regressor algorithm’s flow chart.

### 3.3. Model Evaluation

The data set was divided into a training group (25%) and a testing group (75%). The training group was used to train machine learning algorithms and create predictive models for maximum consumption data. The testing group was used to evaluate the performance of these models. This process is illustrated in Figure 7.

The training and testing process involved a simple partitioning of data to prevent overfitting. Machine learning algorithms’ predictive models were evaluated for performance...
and accuracy using metrics like R2, MSE, MAE, RMSE, and MAPE. Each measurement definition is mentioned in Table 2.

The present research used MSE because of its sensitivity to errors, differentiability, and simplicity of interpretation. The use of RMSE is preferable to MSE because it yields a more easily understandable outcome in the original units of the dependent variable, facilitating straightforward comparison across data sets or models. The mean absolute error (MAE) is a suitable metric where the quantity of errors is more significant than the specific direction of the mistakes, offering a clear and direct evaluation of the model’s performance, and MAPE is particularly valuable for comparing a model’s prediction accuracy to the scale of the actual values.
4. Results and Discussion

The experiment results were reviewed in sections, discussing the initial processing and imputation of missing data, energy consumption prediction for each building, and performance comparisons for random forest, long short-term memory, and gradient boosting regressor models. The presentation of results follows a hierarchy, starting with the normality test, then data preprocessing, and finally model evaluation.

4.1. Normality Testing of Data. The evaluation is aimed at examining the impact of data shape on predictive model development performance, using measures of skewness and kurtosis. Results were compiled in Table 3 to evaluate the data’s shape and potential deviations from normal distribution. To evaluate the normality of the energy demand data, the two values were computed using the aggregated data from each building spanning from January 2020 to January 2023. Figure 8 also depicts the format of the data set for a graphical examination of normality.

Based on Table 3, the data sets for the CLAS, NHAI, and Cronkite buildings were approximately symmetrical and skewed with bidirectional shape distribution. However, there were some differences in the skewness values for each building. The CLAS building showed normal asymmetry due to power consumption and KWS, with a slightly negative skewness indicating a longer left tail. The CHWTON distribution was skewed, with a skewness of 427578, indicating a longer left tail. The nursing and health innovation building had a pronounced asymmetry, with power consumption having a positive skewness and KWS and CHWTON having a negative skewness, indicating balanced tails. The Cronkite building had positive skewness values, indicating a moderate right-skewed distribution. Overall, all three data sets were approximately symmetric, skewed, and bimodal in their form density.

The kurtosis values of all three buildings in Table 3 were less than 0, indicating that their distributions were

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Description</th>
<th>Math form</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R^2 )-squared [37]</td>
<td>The coefficient of determination is used to determine how much of the variance in the dependent variable can be explained by the independent variables.</td>
<td>( R^2 = 1 - \frac{SS_{res}}{SS_{tot}} )</td>
</tr>
<tr>
<td>Mean squared error [38]</td>
<td>A regression metric used to calculate the average squared difference between predicted and actual values.</td>
<td>( MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 )</td>
</tr>
<tr>
<td>Root mean squared error [39]</td>
<td>RMSE is a widely used measure for estimating the average variance between predicted and real values in regression tasks.</td>
<td>( RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} )</td>
</tr>
<tr>
<td>Mean absolute error [40]</td>
<td>A regression statistic used to calculate the average absolute difference between predicted and actual values, ignoring the direction of mistakes.</td>
<td>( MAE = \frac{1}{n} \sum_{i=1}^{n}</td>
</tr>
<tr>
<td>Mean absolute percentage error [39]</td>
<td>A commonly used method for determining forecasting error, as it measures the average absolute percent inaccuracy for each time period less actual values divided by actual values, making understanding it simpler due to its scaled units.</td>
<td>( MAPE = \frac{1}{n} \sum_{i=1}^{n} \left</td>
</tr>
</tbody>
</table>
platykurtic. This was also evident in Figure 8, where the probability distribution plot had a higher tail and a larger peak center. However, the Cronkite building had a kurtosis value greater than 0, indicating a leptokurtic distribution with higher variance. CLAS and NHAI had roughly normal distributions, but CLAS had a lower mean than the median. Department CLAS also had an almost normal distribution but with higher skewness and kurtosis. The CHWTON data set had a higher variation compared to the other data sets.

<table>
<thead>
<tr>
<th>Application area</th>
<th>Building name</th>
<th>Power consumption</th>
<th>KWS</th>
<th>CHWTON</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skewness</td>
<td>CLAS</td>
<td>-0.10206</td>
<td>-1</td>
<td>0.42329</td>
</tr>
<tr>
<td></td>
<td>NHAI</td>
<td>0.017914</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>Cronkite</td>
<td>0.805914</td>
<td>-1</td>
<td>0.494</td>
</tr>
<tr>
<td></td>
<td>CLAS</td>
<td>-0.333548</td>
<td>-2.0</td>
<td>-1.019492</td>
</tr>
<tr>
<td></td>
<td>NHAI</td>
<td>-0.777519</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td></td>
<td>CRONKIT</td>
<td>2.620576</td>
<td>2.056333</td>
<td>-0.687182</td>
</tr>
</tbody>
</table>

Figure 8: Probability density for buildings CLAS, NHAI, and Cronkite.
4.2. Data Preprocessing. Based on Figure 9, the original data set had various scale ranges for power consumption factors like KWS, CHWTON, voltage, and building occupants. To verify the prediction capacity of 29 features, multiple approaches like correlation analysis, ensemble analysis, and tree-based models were used. After testing the mentioned methods, the most ideal qualities for projecting energy demand and consumption are as follows:

(1) Previous consumption patterns

(2) Calendar: weekday, month, and season

(3) Demography: A building’s population might influence consumption patterns

(4) Geographical factors such as climate. People will use more electrical appliances at hot and low temperatures, respectively

The study on missing data utilized the missingness matrix to quantify the extent of missing data and identify rows that contained missing values. Upon analyzing Figure 10, it is noteworthy that none of the three data sets exhibited any missing data.
4.3. Feature Selection. Selecting the most crucial features plays a vital role in enhancing the effectiveness, stability, and scalability of our prediction model. Through the utilization of a feature importance assessment method, as summarized in Table 4, we identified the top five influential features: KW, KWS, CHWTON, total houses, and CHWTONgasgas. The ranking of these features is illustrated in Figure 11, which shows the order of their importance. Although the initial analysis considered all 29 parameters, the figure only highlights features that significantly contribute to precision, ensuring a streamlined and informative depiction.

The study is aimed at predicting energy consumption in three educational buildings by identifying key parameters. Through feature selection, we have identified key parameters that significantly impact energy usage. These include “CHwton” or chilled water tons which measures the cooling capacity of chilled water systems, representing the heat energy required to melt one ton of ice in 24 hours. Additionally, “KW” denotes the power consumption of electrical equipment and lighting systems within the buildings. “Total-lightbulb” denotes the aggregate number of light bulbs or lamps within the buildings, crucial for various assessments. Furthermore, aspects of HVAC systems, like “CHWTON-galsgas,” offer insights into chilled water and gas usage. Moreover, “Combined mmBTU” measures the heat required to raise the temperature of water by one degree Fahrenheit. The feature selection process helps identify the most influential parameters for the predictive model, enabling more accurate energy consumption forecasts.

4.4. Performance Evaluation and Comparison. The prediction models’ performance was evaluated by comparing multiple methods for each building after training and testing. Comparative results are shown in Table 5.

Based on the performance evaluation measurements presented in Table 5, the GBR method exhibited outstanding performance across all buildings. Notably, the determination coefficients were remarkably high, reaching 0.998 for Cronkite, 0.984 for CLAS, and 0.845 for NHAI. Furthermore, the corresponding mean squared error (MSE) values were 8.148, 5.09, and 9.17, respectively. The root mean squared error (RMSE) and mean absolute error (MAE) also supported these results, indicating that GBR outperformed other methods and yielded the best values. Additionally, when assessing the mean absolute percentage error (MAPE) results, GBR surpassed the other methods, demonstrating the lowest error percentage. The LSTM method exhibited lower determination coefficients compared to the GBR results, with values of 0.86 for CLAS, 0.7772 for NHAI, and 0.7609 for Cronkite. However, when comparing LSTM to the RF method, the performance varied across buildings.
Specifically, in the Cronkite building, the random forest method outperformed LSTM with an R2 value of 0.89. Nevertheless, in terms of other metrics such as MSE and RMSE, LSTM yielded comparably smaller values than the RF method. Moreover, there was a significant difference in the MAPE results, with LSTM generating fewer errors compared to random forest. This observation suggests that, in terms of errors, LSTM performed better and produced a lower number of errors compared to RF. According to the forecast evaluation, the square error method was deemed a more suitable evaluation metric for assessing the accuracy of the predictions. Following this examination, it became clear that the gradient boosting regressor (GBR) method performed the best across all buildings.

Considering the data presented in Table 6, it is evident that the algorithm closest to the real testing values is the gradient boosting regressor, demonstrating good precision.
long short-term memory (LSTM) method follows in second place, and the random forest algorithm comes last in terms of accuracy in predicting average consumption. In the context of result validation, K-fold cross-validation is a highly suitable technique for our case due to its inherent advantages. By partitioning the data set into K subsets, each containing a representative sample of the data, K-fold cross-validation ensures thorough training and validation of the model. This approach maximizes data utilization and minimizes bias, as every data point is utilized for both training and validation across different folds. Furthermore, the averaging of performance metrics over multiple splits provides a robust evaluation, effectively reducing the variance associated with a single train-test split. Additionally, K-fold cross-validation facilitates better generalization by assessing the model’s performance across diverse subsets of the data, ensuring that it can effectively handle various scenarios. Its utility extends to hyperparameter tuning, enabling the comparison of different parameter configurations across multiple validation sets.

In our scenario, we choose 5-fold cross-validation for its moderate data set size, balancing computational efficiency and robust performance estimation. This method ensures reliable model evaluation without excessive computational overhead and aligns with common practices in the field, allowing easier comparison with existing literature and benchmarks. Table 7 provides the outcome of the 5-fold cross-validation.

A line graph comparison was used to better demonstrate the difference between the actual and anticipated average consumption levels, as depicted in Figures 12–14. In addition, Figures 15–17 show the graphical presentation of the regression line for the three buildings. In the CLAS and Cronkite buildings, the gradient boosting regressor (GBR) produces a symmetric regression line, indicating that its predicted values closely align with the actual ones. Conversely, for the NHAI building data set, characterized by nonsymmetrical data, long short-term memory (LSTM) outperforms other models due to its ability to capture temporal dependencies.

However, in the case of NHAI, the performance difference between LSTM and GBR is minimal, highlighting the suitability of both algorithms for different data characteristics. GBR excels in all cases, while LSTM’s recurrent nature makes it valuable for handling nonlinear, time-dependent data.
From the analysis of all the tables and figures, we conclude that the best performances are consistently achieved by the gradient boosting regressor (GBR). GBR’s sequential training approach trains weak learners sequentially, correcting errors from previous iterations, and fine-tuning the model’s predictive capabilities with each step. Additionally, gradient descent optimization minimizes prediction errors, leading to more accurate predictions. Following GBR, long short-term memory (LSTM) stands out as it is specifically designed for handling sequential data, making it well-suited for time series forecasting and similar tasks. Its ability to understand and process temporal patterns contributes to accurate predictions in time-dependent scenarios. Lastly, the random forest algorithm also delivers good results, particularly when it comes to capturing complex nonlinear correlations between features and the subject variable, and its ability to model complex interactions and patterns makes it effective.

The CLAS building has a significantly higher energy consumption rate, exceeding 30 kWh, in contrast to the other buildings. The main reason for this difference is the large surface area and the simultaneous use for many educational objectives. On the other hand, the Cronkite building has an energy consumption rate of 26 kWh/h, while NHA1 has a consumption rate of 12 kWh per hour. Predictive modeling approaches are necessary for efficient energy allocation and management. Within this particular instance, the gradient boosting regressor model demonstrates its superiority in effectively predicting outcomes for both the CLAS and Cronkite buildings. The choice is backed by the model’s remarkable performance metrics, as shown by its coefficient of determination ($R^2$-squared) values of 0.99 for Cronkite and 0.98 for CLAS. This model improves the accuracy of forecasting by offering proactive insights into the energy needs of each building. It also helps in preventing energy loss before it happens and promotes efforts to reduce energy usage.

5. Comparison with the Previous Study

The study compared three algorithms: random forest, LSTM, and gradient boosting regressor, revealing their performance in forecasting monthly average consumption.
The development of prediction models demonstrated their capabilities, urging further optimization. The findings also led to a comparative analysis with previous machine learning studies. In the first research conducted by Khaoula et al. in 2022 [40], four machine learning algorithms were implemented to predict energy demand for a commercial building over two years. The algorithms used were multiple linear regression (MLR), long short-term memory (LSTM), simple linear regression (LR), and random forest (RF). The results indicated that LSTM performed the best, followed by RF, MLR, and LR, providing valuable insights into the regression algorithms’ capabilities. In the second research, Khaoula et al. in 2023 [41] examined energy consumption prediction in a low-energy house over four months. Unlike the first research, this time, the prediction considered not only the house’s energy but also its appliances. Three machine learning algorithms, namely, artificial neural networks (ANN), recurrent neural networks (RNN), and random forest (RF), were employed for tests. Recurrent neural networks especially LSTM once again outperformed the other algorithms, achieving an impressive accuracy of 96%. RF was followed with 88% accuracy. However, ANN yielded negative predictions, indicating its unsuitability for time series data sets. Furthermore, in their research, Khaoula et al. [42] used three deep learning algorithms—recurrent neural networks (RNNs), artificial neural networks (ANNs), and autoregressive neural networks (AR-NNs)—to forecast the total load of HVAC systems. The results showed that the autoregressive neural network model outperformed the other two due to its ability to capture temporal dependencies and patterns in time series data, which is crucial for HVAC load prediction. AR-NNs use a simpler architecture, focusing on past observations to predict future values, and their autoregressive nature allows them to effectively model the self-dependence of time series data, leading to more accurate predictions.

Drawing insights from these three studies, significant findings emerge regarding the efficacy of regression algorithms for energy consumption prediction. Specifically, long short-term memory (LSTM) and random forest (RF) consistently emerge as top performers, especially in handling time series data. However, our research introduces a novel aspect by exploring the effectiveness of gradient boosting regressor (GBR), which yielded exceptional results. Notably, GBR
achieved remarkable precision, boasting an impressive accuracy of 98.9%. Moreover, compared to other algorithms, GBR demonstrated superior performance with fewer errors, as evidenced by lower root mean square (RMS), mean absolute error (MAE), and mean absolute percentage error (MAPE) values. This underscores the potential of GBR as a formidable contender in energy consumption prediction tasks, offering a promising alternative to LSTM and RF in certain contexts.

6. Perspectives and Future Work

For future contributions, we plan to optimize the GBR model by increasing the data used for training and prediction, which may improve efficiency and performance on larger data sets. We intend also to apply a novel approach to the gradient boosting optimizer to fine-tune the model’s parameters and hyperparameters more effectively. These efforts are aimed at enhancing the GBR algorithm’s performance for accurate energy consumption forecasting and other applications.

Another significant contribution of our future research lies in the utilization of transformer models for predicting diurnal energy consumption patterns. Transformers, originally designed for natural language processing tasks, have shown remarkable capabilities in capturing long-range dependencies in sequential data, making them well-suited for time series forecasting tasks as well. By applying transformer architectures to predict diurnal energy consumption, we aim to leverage their ability to effectively model complex
temporal patterns and dependencies inherent in energy consumption data. Our case study focuses on commercial and institutional buildings, where accurate energy consumption prediction is crucial for optimizing building operations, reducing costs, and minimizing environmental impact.

7. Conclusion

Our major focus in this research is developing an energy consumption forecasting model given the environment of three institutional buildings that have adopted the smart building ecosystem. From January 2020 to January 2023, the collected energy consumption data was subjected to statistical analysis to assess its normality. The skewness and kurtosis values showed that the data had a variety of distribution characteristics.

The predictive model development process involved data preprocessing, which included handling missing data and identifying feature importance. For this research’s objective, three supervised machine learning methods, namely, gradient boosting regressor (GBR), long short-term memory (LSTM), and random forest (RF), were selected as the algorithms for the predictive model. The comparison of these strategies was based on an assessment of their production structures and prediction abilities. The results of our model training and testing indicated that each strategy performed differently for each building. Remarkably, the GBR approach continually produced the most promising outcomes, cementing its position as the best-performing strategy across all three buildings: CLAS, NHAI, and Cronkite. GBR’s mean absolute percentage error (MAPE) values were 9.337, 12.338, and 4.045 for CLAS, NHAI, and Cronkite, respectively. Additionally, GBR achieved a lower mean absolute error (MAE) for CLAS and Cronkite (71.04 and 53.77, respectively), while RF and LSTM yielded lower MAE results for these two buildings. Moreover, while computing average consumption using demand data, it was shown that the gradient boosting regressor (GBR) displayed greater accuracy in anticipating demand. This performance outperformed all other approaches in all buildings.

In terms of future study recommendations, it is suggested to use more powerful computers or platforms to run the LSTM algorithm, potentially improving its performance. Additionally, exploring hybrid or ensemble methods may be beneficial, as they have shown higher accuracy than single regressors. Lastly, a comparison with another smart building could be included to distinguish and validate the obtained results. These recommendations can further enhance the understanding and applicability of the energy consumption predictive model.

Data Availability

The collected data was saved in an open-source website server [43] and could be manually downloaded from the platform’s website in the form of a CSV file with any sort of aggregation [43] (https://portal.emcs.cornell.edu/d/2/dashboard-list?orgId=2).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors thank the National Center for Scientific and Technical Research (CNRST) for supporting and funding this research.

References

environment accuracy, technology using machine learning model with high extrapolation accuracy,

for energy use of appliances prediction, sumption in campus buildings using long short-term memory networks,


"Coefficient of determination (R^2)," Encyclopedia of Statistics in Quality and Reliability, 2008.


"Evaluation and comparison of energy consumption prediction models," in International Conference on Advanced Technologies for Humanity, Marrakech, Morocco, 2022.


"Real time utility use data," https://portal.emcs.cornell.edu/d/2/dashboard-list?orgId=2.